Neuroscience and Biobehavioral Reviews 160 (2024) 105640

Contents lists available at ScienceDirect

Neuroscience and Biobehavioral Reviews

journal homepage: www.elsevier.com/locate/neubiorev

Predicting treatment outcome based on resting-state functional connectivity in internalizing mental disorders: A systematic review and meta-analysis

Charlotte Meinke [a,*], Ulrike Lueken [a,b], Henrik Walter [c,1], Kevin Hilbert [a,d,1]

[a] Department of Psychology, Humboldt-Universität zu Berlin, Germany
[b] German Center for Mental Health (DZPG), partner site Berlin/Potsdam, Germany
[c] Charité Universtätsmedizin Berlin, corporate member of FU Berlin and Humboldt Universität zu Berlin, Department of Psychiatrie and Psychotherapy, CCM, Germany
[d] Department of Psychology, Health and Medical University Erfurt, Germany

ARTICLE INFO

Keywords:
Treatment outcome
Depression
Post-traumatic stress disorder
Machine learning
Prediction
Feature importance
Resting-state
Functional connectivity

ABSTRACT

Predicting treatment outcome in internalizing mental disorders prior to treatment initiation is pivotal for precision mental healthcare. In this regard, resting-state functional connectivity (rs-FC) and machine learning have often shown promising prediction accuracies. This systematic review and meta-analysis evaluates these studies, considering their risk of bias through the Prediction Model Study Risk of Bias Assessment Tool (PROBAST). We examined the predictive performance of features derived from rs-FC, identified features with the highest predictive value, and assessed the employed machine learning pipelines. We searched the electronic databases Scopus, PubMed and PsycINFO on the 12th of December 2022, which resulted in 13 included studies. The mean balanced accuracy for predicting treatment outcome was 77% (95% CI: [72%- 83%]). rs-FC of the dorsolateral prefrontal cortex had high predictive value in most studies. However, a high risk of bias was identified in all studies, compromising interpretability. Methodological recommendations are provided based on a comprehensive exploration of the studies' machine learning pipelines, and potential fruitful developments are discussed.

1. Introduction

Internalizing mental disorders including depressive disorders, anxiety disorders, obsessive compulsive disorders, and post-traumatic stress disorder are highly debilitating, ranking among the top ten causes for global years lived with disability (GBD, 2019 Mental Disorders Collaborators, 2022), and are associated with a substantial reduction of quality of life (Mack et al., 2015). These disorders are often grouped together (e.g., Hettema et al., 2006; Wergeland et al., 2021) because their symptoms have shown to load on a shared latent factor, commonly referred to as the internalizing factor (e.g., Andrews, 2018; Kotov et al., 2017). This factor is mainly characterized by distress and fear and also underlies their high comorbidity (Kessler et al., 2011).

Last decades of research have yielded effective treatments for these disorders, including psychotherapy, pharmacotherapy, electroconvulsive treatment (ECT), and repetitive transcranial magnetic stimulation (rTMS; e.g., see meta-analyses of Carpenter et al., 2018; Cuijpers et al., 2013; Dalhuisen et al., 2022; Mutz et al., 2019). However, each of these treatments comes with a high proportion of patients whose condition does not improve after treatment (non-responders, e.g., see reviews of Fitzgerald, 2020; Fonseka et al., 2018; Loerinc et al., 2015; Papakostas and Fava, 2009). These high rates of non-responders across treatments may indicate that there is no one-size fits all treatment and suitability varies among individuals or subgroups of patients. Following the concept of precision mental healthcare (DeRubeis, 2019), allocating patients a priori to the treatment most promising for them could reduce non-response rates. A necessary condition for such a treatment allocation is a sufficiently accurate prediction of treatment outcome on a single-subject level.

From a methodological standpoint, machine learning approaches are particularly well-suited for this endeavor. In contrast to conventional statistical modeling, which predominantly aims at explaining existing

* Correspondence to: Humboldt-Universität zu Berlin, Faculty of Life Sciences, Department of Psychology, Unter der Linden 6, Berlin 10099, Germany.
E-mail addresses: charlotte.meinke@hu-berlin.de (C. Meinke), ulrike.lueken@hu-berlin.de (U. Lueken), henrik.walter@charite.de (H. Walter), kevin.hilbert@hu-berlin.de (K. Hilbert).
[1] Shared authorship.

https://doi.org/10.1016/j.neubiorev.2024.105640
Received 29 June 2023; Received in revised form 29 February 2024; Accepted 21 March 2024
Available online 26 March 2024

data, the core objective of machine learning is the accurate prediction of new data (Sidey-Gibbons and Sidey-Gibbons, 2019; Yarkoni and Westfall, 2017). This shift in focus gives rise to two central distinctions between statistical modeling and machine learning: the assessment of the model's performance and the models or algorithms employed. First, both approaches substantially diverge in their criteria and their procedure for determining a well-performing model. In statistical modeling, a well-performing model is one that effectively explains the data (e.g., a logistic regression model with a high R-squared indicating goodness-of-fit). On the other hand, in machine learning, a well-performing model is one that discriminates effectively between two or more classes in new data (e.g., achieving high predictive accuracy). Hence, instead of evaluating a model's performance in the data set on which it has been trained, machine learning approaches apply the fitted model to ideally new data and assess its performance there. Since entirely new data are often unavailable, cross-validation techniques have been developed, iteratively dividing the dataset into a training set for model fitting and a test set for model evaluation. Several metrics to evaluate a model's classification performance on the test set(s) exist, combining mainly the number of correctly and falsely predicted cases. One of the most general and most frequently used metrics is accuracy, summarizing the proportion of correctly classified (positive and negative) cases in relation to the total number of cases. However, its interpretability diminishes when being based on imbalanced classes (e.g. 60% nonresponders, 40% responders), a factor overlooked in several studies. In such cases, other metrics are recommended, including the Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC) and the balanced accuracy (see Thölke et al., 2023 for a more detailed discussion).

The second central distinction between machine learning and statistical modelling are the models or algorithms employed. Seeking to explain the observed data in a clear manner, statistical models often determine possible relationships between variables and employ a small number of dependent variables. For instance, logistic and linear regression assume a linear relationship between independent and dependent variables, with interaction effects only considered when explicitly added to the model. In contrast, machine learning approaches include a diverse array of algorithms capable of effectively handling numerous variables and capturing nonlinear relationships. Some of the most common algorithms are support vector machines (SVM), random forests and neural networks. To conclude, machine learning approaches are particularly well-suited for testing the possibility of pre-treatment prediction of treatment outcomes, given their inherent design for this application, validating models by their performance on unseen data, and employing versatile algorithms (see Sidey-Gibbons and Sidey-Gibbons, 2019 for a more in-depth introduction to machine learning).

A large number of studies has employed machine learning approaches to predict treatment outcome in internalizing mental disorders, using a wide range of modalities including demographic, clinical, EEG, and (f)MRI data (e.g., see reviews of Cohen et al., 2021; Karvelis et al., 2022; Vieira et al., 2022). One promising subtype of fMRI data is resting-state functional connectivity, relying on BOLD (blood--oxygenation level dependent) contrast imaging to measure the local blood supply in the brain as a proxy of neural activity. The basis for calculating resting-state functional connectivity is a resting-state scan, in which participants are directed to remain motionless for approximately 8–10 minutes without engaging in any specific task or receiving visual stimuli. Based on these data, functional connectivity between grey matter brain regions is calculated as their statistical correlation of BOLD signals over time (for an easy introduction into resting-state functional connectivity, see Lv et al., 2018). Neuroimaging data and more specifically resting-state functional connectivity appeared to be particularly valuable in previous reviews and meta-analyses comparing prediction performances across different modalities (Del Fabro et al., 2023; Vieira et al., 2022). Moreover, resting-state functional connectivity seems promising due to its shared alterations across internalizing disorders, including disorder-specific variations (e.g., Williams, 2017). The size and type of these alterations may impact treatment response likelihood, independent of the type of treatment. Additionally, compared to other neuroimaging modalities, resting-state data can be assessed relatively consistently across sites and with a low burden on patients, facilitating the generation of larger samples, as required for machine learning.

However, the combination of resting-state functional connectivity and machine learning goes along with several methodological challenges. One of the biggest challenges is the high dimensionality of resting-state functional connectivity data (Khosla et al., 2019). In theory, an extensive number of functional connectivities can be computed from a resting-state scan in typical resolution. A normalized brain scan with an isotropic voxel size of 2 mm has around 124.000 voxels of grey matter. Thus, when calculating functional connectivity between all voxels, more than 15 billion (124.000 ×123.999) functional connections would be initially available to predict treatment outcome. Such a large number of predictive variables (in machine learning called features) cannot be handled by current machine learning classifiers, especially not with sample sizes of few hundred patients, which is usually the upper limit for longitudinal interventional studies (also known as the curse of dimensionality or small-n-large-p-problem; Mwangi et al., 2014).

The current status of addressing challenges specific to resting-state functional connectivity has not been summarized, as previous reviews have rather examined the predictive ability across different neuroimaging modalities in general (Cohen et al., 2021; J. Lee et al., 2022; Y. Lee et al., 2018; Vieira et al., 2022). Furthermore, the majority of these reviews lacked a comprehensive quality control that thoroughly examined the employed machine learning approach (Cohen et al., 2021; J. Lee et al., 2022; Y. Lee et al., 2018). The reviews of Y. Lee et al. (2018) and Cohen et al. (2021) did address the overall risk of bias to some extent by investigating publication bias. However, Y. Lee et al. (2018) did not conduct any assessment of study quality. In contrast, Cohen et al. (2021) evaluated study quality, including risk of bias, with the QUADAS-2 tool, but overlooked the bias introduced by the design of the machine learning pipeline, as the tool primarily focuses on the validation of diagnostic tests. This lack of attention to the risk of bias introduced by the design of the machine learning pipeline is problematic, considering the common occurrence of inappropriate applications of machine learning in the field (Meehan et al., 2022). Finally, existing reviews did not systematically assess which features contributed to a successful prediction.

To fill these gaps, the aim of this systematic review was three-fold: First, to examine how well treatment outcomes in internalizing disorders can be predicted by features based on resting-state functional connectivity (research question 1 = RQ 1), taking into account the studies' risk of bias using the Prediction Model Study Risk of Bias Assessment Tool (PROBAST). Second, to assess how features with high predictive value were identified (RQ 2.1) and which features were particularly important for the prediction (RQ 2.2). Third, to provide an overview of how the different studies addressed the curse of dimensionality, i.e. how they reduced the large number of theoretically initially available functional connectivities to a small set of features to be used in the final classifier(s) (RQ 3). By addressing these questions, we aimed to give a realistic estimate of the potential of machine learning and resting-state functional connectivity (RQ 1), to assist future researchers in making informed methodological decisions by summarizing current practices and identifying methodological shortcomings (RQ 2.1, RQ 3), and to provide evidence for an a priori selection of brain regions which might be relevant for predicting treatment outcome in future machine learning studies (RQ 2.2).

2. Methods

2.1. Search strategy and inclusion criteria

The electronic databases Scopus, PubMed and PsycINFO were

searched for relevant studies from inception up the 12th of December 2022. Search terms encompassed keywords for resting state, primary disorder, treatment, and machine learning (see supplement S1 for the specific search terms in all databases). Additionally, reference lists of eligible studies and review articles were screened. The following inclusion criteria were applied: 1) publication in a peer-reviewed journal, written in English, 2) analyzing a sample of patients with one of the following disorders as primary disorder: unipolar depressive disorders, anxiety disorders, obsessive compulsive disorder, or post-traumatic stress disorder, 3) predicting outcome to any treatment (behavioral, pharmacological, placebo, or neuroscience-informed) that aimed to improve the patients` condition 4) using a machine learning approach 5) predicting treatment outcome as a categorical outcome 6) reporting at least one classifier whose input features are exclusively based on resting-state functional connectivity. Anticipating a limited number of studies meeting our inclusion criteria, we refrained from delineating any additional criteria beyond those specified. After an initial abstract screening by the first author (CM), all remaining studies were submitted to a full-text screening, independently performed by two authors (CM, KH). Disagreements were resolved by discussion. The review part has been preregistered with PROSPERO (CRD42022370949). The meta-analytic summary of classification accuracies was not part of the preregistration as it was conducted in response to the reviewers` recommendations.

2.2. Data extraction

The following data, ordered by research questions, were extracted. Study characteristics and RQ 1: first author, year, primary disorder, age group, treatment, definition of response and/or remission, sample size, way of estimating the underlying functional connectivities, type of functional-connectivity-based input features, algorithm(s) of the final classifier(s), validation method, classification metrics of the best model reported. RQ 2: way of detecting features with high predictive value, level of resolution of investigating high predictive values, features that showed high predictive value. RQ 3: approaches to reduce the number of input features. Data extraction was performed by CM and checked by KH. The original table of data extraction as well as a R-script for reproducing all our analyses and plots in R (R Core Team, 2020) can be found here: https://osf.io/y69ke/.

2.3. Risk of bias assessment

The best model of each study was assessed for risk of bias using PROBAST (Prediction model study risk of bias assessment tool; Wolff et al., 2019), a tool developed for predictive modelling in healthcare. Based on 20 signaling questions, each model was judged as having low or high risk of bias, in each of four domains (participants, predictors, outcome, analysis) and in total.

2.4. Data synthesis

2.4.1. RQ 1 Meta-analysis on balanced classification accuracies

To answer RQ 1 (How well can treatment outcome in internalizing disorders be predicted by features based on resting-state functional connectivities?), we estimated the mean balanced classification accuracy in a meta-analysis using the classification accuracy of each study's best model. We focused on each study's best model as sufficient performance metrics were mostly only reported for those. This has been a common procedure in systematic reviews and/or meta-analyses on machine learning (e.g., Bondi et al., 2023; Vieira et al., 2022). We chose accuracy instead of other metrics such as precision, recall/sensitivity, or specificity, which focus on the prediction of one of two classes (either response or nonresponse), for two reasons. First, given that current models are far from any clinical application, it is unclear whether predicting one class is more crucial than the other. Therefore, prioritizing the evaluation of overall model performance, summarized by accuracy, seemed most pertinent. Secondly, opting for accuracy was more practicable, as it was the only metric consistently reported across all studies. In contrast, sensitivity and specificity, the most frequently reported metrics among those focusing on the prediction of one of the two classes, were absent in 4 out of the 13 studies reviewed. In addition, aggregating these metrics across studies would have been challenging, as it was not always apparent to which class the metrics referred. For instance, in some studies it was unclear whether specificity described the ability to predict response or nonresponse.

However, using the classification accuracy has the disadvantage that its meaningfulness diminishes when classes are imbalanced, as accuracies above 50% can easily reached by a model that is systematically predicting the more frequent class (Thölke et al., 2023). For instance, consider a binary classification scenario where nonresponders constitute 70% of the cases. In this context, a classification accuracy of 70% might not truly signify high predictive performance. Instead, it could be attributed to a model lacking genuine predictive ability, merely predicting a nonresponder status for all cases. Some studies with imbalanced classes took this into account by reporting the balanced accuracy or additional evaluation metrics. The balanced accuracy is commonly calculated as the mean of sensitivity and specificity (e.g., Brodersen et al., 2010). Hence, when reported, we calculated missing balanced accuracy values based on these metrics. However, this could not be done for all studies with imbalanced classes. Therefore, we estimated a proxy of balanced accuracy for those remaining studies, using the following formula: Proxy of balanced accuracy = (raw accuracy – relative frequency of the more frequent class) + 50%. This proxy is based on the idea that the accuracy achieved by a dummy classifier always predicting the more prevalent class (= the relative frequency of the more frequent class) represents the chance-level. The improvement above chance-level is thus calculated by subtracting the chance-level from the raw accuracy. To get the final proxy of balanced accuracy, it is added to a chance-level of 50%, as it would exist when classes are balanced. This formula is not a prevalent method in machine learning for taking class imbalances into account, as more suitable metrics such as balanced accuracy and AUC exist when evaluating a model on original data. Its relevance only emerges when summarizing accuracy values across studies and other performance metrics controlling for class imbalance are lacking.

Similar to previous meta-analyses (Y. Lee et al., 2018; Vieira et al., 2022), we conducted a meta-analysis for proportions, treating accuracy values as proportions of correctly classified cases. The R-package meta was employed for these analyses (Balduzzi et al., 2019). As we anticipated a considerable heterogeneity in classification accuracy between studies, we fitted a random effect model and applied Knapp-Hartung adjustments (Knapp and Hartung, 2003) to calculate the confidence interval around the mean estimated accuracy. All analyses were conducted using Freeman Tukey double arcsine-transformed proportions to stabilize error variances (Barendregt et al., 2013). Between-group heterogeneity in the random effect model was estimated with the restricted maximum likelihood estimator (Viechtbauer, 2005). The mean estimated accuracy under the random effect model was calculated by pooling studies accuracies by the inverse of their error variance. Individual study confidence intervals were calculated using the Clopper-Pearson (i.e., exact binomial interval) method. Heterogeneity between studies was evaluated using the I^2 statistic and the Cochrane's Q-test. We interpreted I^2 following the recommendations from Higgins and Thompson (2002), with 25%, 50%, and 75% as low, moderate, and high, respectively. To explore potential sources of heterogeneity, we performed subgroup analyses for treatment, diagnosis, and risk of data leakage (PROBAST question 4.8) as well as a meta-regression on sample size.

2.4.2. RQ 2 Features with high predictive value

To answer RQ 2.1 (Which approaches are taken to draw inferences about the predictive value of specific features?), we extracted approaches that were used to assess which feature (sets) contributed or led

to a prediction above chance-level and then clustered them qualitatively into suitable groups. To account for the expected heterogeneity of approaches, we used the term "predictive value" instead of "feature importance", as the latter is often used to describe the contribution of a feature in a final classifier, which represents only one of various methods.

To answer RQ 2.2 (Which features have high predictive value for the prediction of treatment outcome?), we applied a five-step procedure. First, we extracted which features were reported to have high predictive value. Then, we examined the level of resolution at which these features were reported. For instance, certain studies indicated that a singular functional connectivity between two brain regions holds predictive value, whereas others reported the entire array of functional connectivities between a particular brain region and a subset of other regions as having high predictive value. Since the majority of studies fell into the second category, we decided to summarize findings across studies at the level of brain regions (instead for example, at the level of single functional connectivities). Thus, in a third step, we collected those brain regions whose functional connectivity had high predictive value. Fourth, we grouped these brain regions into larger areas to provide a more comprehensive overview, based on the 22-regions Human Connectome Project multimodal brain parcellation (Glasser et al., 2016; Huang et al., 2022), common subcortical areas, and findings from previous literature. This approach resulted in the following coarse brain areas: visual areas, sensorimotor areas, inferior temporal gyrus, middle temporal gyrus, superior temporal gyrus, parahippocampal gyrus, superior parietal lobule, inferior parietal lobule, posterior cingulate cortex, anterior cingulate cortex (ACC), medial prefrontal cortex (PFC), precuneus, orbitofrontal cortex, ventrolateral PFC, dorsolateral PFC (DLPFC), amygdala, hippocampus, insula, basal ganglia, and thalamus. The Glasser's 22-regions parcellation served solely as inspiration for categorizing brain regions. We did not align coordinates between brain regions and the Glasser parcellation; instead, we relied on the labels provided by the studies for grouping. However, it is important to highlight three key distinctions from the Glasser's parcellation in our approach: First, we merged Glasser's visual areas 1–5 into one visual area, as this parcellation seemed too fine-grained for our endeavor. Second, we splitted Glasser's area 19 (ACC and medial prefrontal cortex) into ACC, medial PFC, and precuneus, as these regions and their role in psychopathology and treatment outcome have been discussed separately. Third, we parcellated the temporal lobe in an anatomical way, because the Glasser's more functionally-based parcellation could not be imposed on our studies' findings.Then, in a fifth step, in order to account for the fact that not all studies employed whole-brain analyses, we assessed for each study and each brain area whether there was an initial opportunity to demonstrate high predictive value. This was for example not the case for all coarse brain areas when the analysis focused solely on functional connectivities within a subset of brain regions or when whole-brain analyses were confined to cortical areas.

2.4.3. RQ 3 Approaches to reduce the number of features

To answer RQ 3 (Which approaches are taken to reduce the large amount of initially available functional connectivities to a small set of features to be used in the final classifier(s)?), the extracted approaches were grouped into suitable categories (here: approaches that served an initial reduction preceding feature generation and approaches that served feature reduction).

3. Results

3.1. Search results and study characteristics

The initial search identified 240 unique records. After screening for eligibility by title and abstract, 49 studies underwent a full text screening. Finally, 13 studies, each using a different sample, were included in the systematic review (see flowchart in Fig. 1).

Even though including a variety of internalizing disorders in the search term, most of the finally selected studies predicted treatment outcome in patients with unipolar depressive disorders (n = 11), 2 studies focused on patients with post-traumatic stress disorders, while

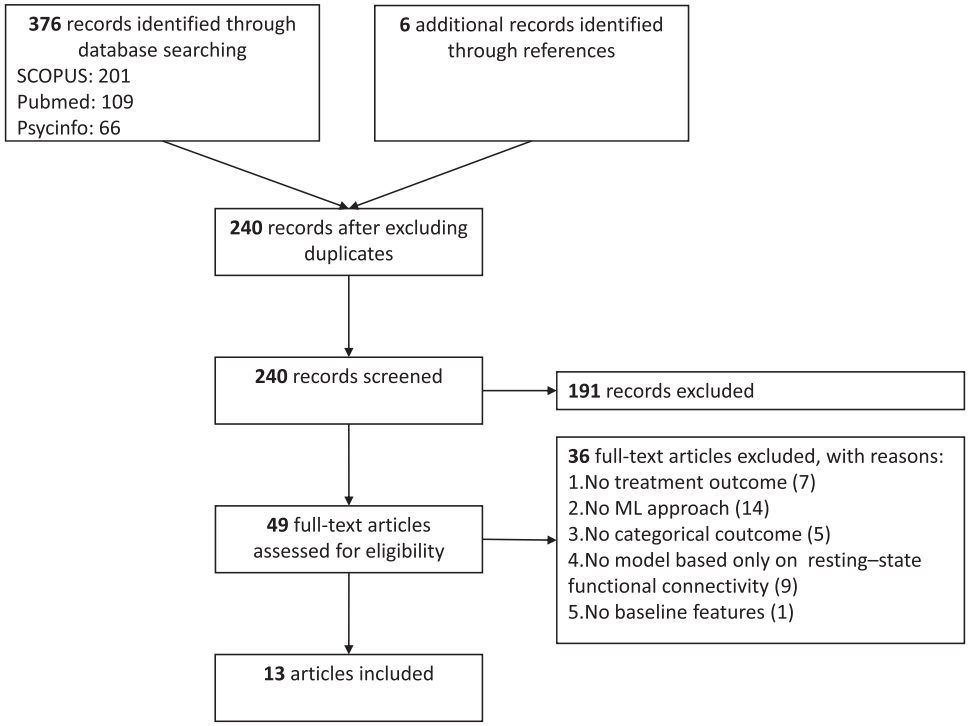

Fig. 1. PRISMA Flowchart. We excluded one study because it did not utilize any baseline features to predict treatment outcomes, even though the explicit criterion "Using only baseline features" was not included in our exclusion criteria. However, we interpreted inclusion criterion 3) "Predicting outcome to any treatment" as encompassing the use of only baseline features, given our understanding that prediction occurs before treatment.

there was no study on patients with obsessive compulsive disorder or anxiety disorders fulfilling the inclusion criteria (see Table 1 for study characteristics). More specifically, all patients with unipolar depressive disorders met the DSM-IV criteria for a major depressive episode, mostly evaluated through (semi)-structured interviews such as the SCID or the MINI. Patients in studies on PTSD satisfied the DSM-IV criteria for PTSD either in full (Zhutovsky et al., 2019) or at least partially (Zhutovsky et al., 2021). Criteria were assessed using PTSD-specific semi-structured interviews such as the CAPS, or, focusing on children and adolescents, the CAPS-CA and the ADIS-P (Zhutovsky et al., 2021). Except for Zhutovsky et al., 2021, all studies exclusively included adult participants. Regarding symptom severity, the majority of patients exhibited moderate symptoms, although certain individual studies specifically included patients with more severe symptoms (Hopman et al., 2021; van Waarde et al., 2015). The treatment patients underwent varied largely across studies: 5 studies employed medication (Harris et al., 2022; Kang and Cho, 2020; Kong et al., 2021; Pei et al., 2020; Tian et al., 2020; H. Wu et al., 2022), 3 studies ECT (Moreno-Ortega et al., 2019; Sun et al., 2020; van Waarde et al., 2015), 2 studies psychotherapy (cognitive behavioral therapy and eye movement desensitization and reprocessing; Zhutovsky et al., 2019; Zhutovsky et al., 2021), 2 studies rTMS (Drysdale et al., 2017; Hopman et al., 2021), and 1 study mixed medication and psychotherapy (Schultz et al., 2018). The diversity in treatments and their duration contributed to a large variability in the timing of the assessment of post treatment outcome, ranging from 7 weeks (Schultz et al., 2018) to 8 months (Zhutovsky et al., 2019). Furthermore, it is noteworthy that two studies specifically concentrated on the early response to medication, measuring and predicting outcomes after a brief period of 2 weeks of treatment (Pei et al., 2020; Tian et al., 2020). Treatment outcomes were always measured in terms of symptom severity, mostly using clinician-rated measures such as the HDRS (HDRS-6, HDRS-17, and HDRS-24) MADRS, and CAPS (More information on the definition on treatment outcome can be found in Table 1).

3.2. RQ 1 Meta-analysis of balanced classification accuracy

The meta-analysis was based on 13 studies with $n = 972$ observations. Aiming to summarize balanced instead of raw classification accuracies, balanced accuracies were calculated from sensitivity and specificity for $n = 5$ studies and estimated using the proposed proxy for $n = 4$ studies. Fig. 2a depicts the difference between raw and balanced accuracies for these studies.

Further details are available in the forest plot displayed in Fig. 3. Our analysis revealed substantial between-study heterogeneity, with an estimated variance (τ^2) of 0.009 (95% CI [0.003; 0.027]). The heterogeneity was large, as indicated by a I^2 of 75% (95% CI [56%, 85%]) and a significant Cochrane's Q-test ($X^2(23) = 47.5$, $p < 0.01$).

In subgroup analyses, neither treatment type, primary disorder, nor the presence of data leakage could account for the observed between-study heterogeneity, potentially due to limited subgroup sizes. Including sample size as a moderator, however, reduced the between study heterogeneity to 57.8%. The meta-regression analysis revealed that lower sample sizes were associated with higher classification accuracies (ß based on transformed proportions = −0.0017, t(11) = −2.5, $p = 0.0280$), explaining around 46% of the observed variance. This relationship is depicted in Fig. 2b. Given the significant between study heterogeneity, we chose not to create a funnel plot and perform asymmetry analyses, following various recommendations in the field (Ioannidis and Trikalinos, 2007; Terrin et al., 2003).

3.3. RQ 2.1 Approaches to evaluate predictive value

We grouped the approaches used to draw inferences on the features´ predictive values into three categories: model comparison, selection frequency in feature selection, and feature importance in final classifier. The majority of studies (n = 6) used model comparison. They built and compared models with different sets of input features to assess which set showed the best model performance and thus had the highest predictive value. The compared feature sets included connectivities of different brain regions (Schultz et al., 2018), different combinations of single connectivities (Hopman et al., 2021; Moreno-Ortega et al., 2019), or subject-specific spatial maps of different independent components (van Waarde et al., 2015; Zhutovsky et al., 2019; Zhutovsky et al., 2021).

The "selection frequency in feature selection" was used by 4 studies to draw inferences on the features` predictive value. As previously described, employing feature selection techniques is a common practice to narrow down the initially available features to a more compact set, which is then utilized for the final classifier. When using an internal cross-validation technique as almost all our studies did, feature selection is applied in each iteration. Thus, features which are selected in most iterations are considered as having high predictive value. The studies here used different techniques of feature selection such as Wilcoxon rank sum test (Drysdale et al., 2017), correlation analysis (Sun et al., 2020), SVM with recursive feature elimination (H. Wu et al., 2022), and univariate feature selection (Zhutovsky et al., 2019).

Feature importances in the final classifier were used by 4 studies to assess the features´ predictive value. The category "feature importance in the final classifier" comprises approaches which used measures of feature importance in the final classifier to investigate the features´ predictive value. In general, most classifiers have model-specific measures of feature importance but there exist also a wide variety of model-agnostic approaches. Here, the measures of feature importance varied across studies with feature weights for SVM (Tian et al., 2020; Zhutovsky et al., 2021), position ranking for SVM with recursive feature elimination (Pei et al., 2020), and feature weights in a spatio-temporal graph convolutional network (Kong et al., 2021). Please note that some of the 12 studies which examined the features´ predictive value applied multiple approaches. More information on the specific approaches and categorization per study can be found in table S1.

3.4. RQ 2.2 Important brain regions

As described above, we first examined the level at which features with high predictive value were reported in order to identify the most suitable level for summarizing findings across studies. Most studies (6/12) reported that that the entire array of functional connectivities of a specific brain region had high predictive value and did not focus on, for example, single connectivities. The remaining studies reported high predictive value on levels that could be easily transferred to a brain region level: three studies reported high predictive value for single connectivities, three other studies reported high predictive value for independent components, that were described in terms of common brain regions. In addition, H. Wu et al. (2022) reported high predictive value for specific emotion regulation networks. Please see table S2 for the categorization per study. It is noteworthy that no study reported high predictive value for common functional connectivity networks like the default mode or salience network. Thus, we decided to summarize features which had high predictive value on a brain region level, applying the procedure described above in the methods section. Fig. 4 provides a summary of both the absolute and relative frequency with which a brain region demonstrated high predictive value. The DLPFC was the region whose connectivities were most frequently predictive across studies, both in terms of absolute and relative numbers. Other important brain regions included sensorimotor areas, visual areas, and the basal ganglia.

3.5. RQ 3 Approaches to reduce the large amount of theoretically initially available functional connectivities

We categorized approaches taken to reduce the theoretically initially large number of available functional connectivities to a more manageable set of features into two levels: those serving an initial reduction before feature generation and those serving feature reduction after

Table 1
Characteristics of the included studies.

Study	Primary disorder	Treatment	Definition treatment outcome	N	Responders/ nonresponders	Estimating FCs	Input features	Algorithm(s) of the final classifier (s)	Validation method	Best Acc	Information on models tested
Drysdale, 2017	MDD	rTMS at dorsomedial cortex	response: ≥ 25% ↓ HDRS-17	124	70/54	Pearson correlation controlled for age	whole-brain between-ROI FCs	linear SVM	LOOCV	78*	no other models tested
Harris, 2022	MDD	SSRI	response: ≥ 50% ↓ MADRS	144	67/77	Pearson correlation, partial correlation, tangent	whole-brain between-ROI FCs	logistic regression, linear SVM, radial kernel SVM, random forest	10-fold CV	58**	total number: 240; varying: parcellation, connectivity estimation, dimensionality reduction and classifiers; accuracies: 39% - 61%
Hopman, 2021	MDD	rTMS at left DLPFC	response: ≥ 50% ↓ MADRS	61	33/28	Pearson correlation	4 specific ROI-to-cluster FCs: subgenual anterior cingulate cortex (sgACC) - frontal pole (l), sgACC - superior parietal lobule (l), sgACC - lateral occipital cortex (l), dorsolateral PFC (l) - central opercular cortex (l)	linear SVM	1-fold V	85**	total number: 14; varying: features; accuracies: ca. 38% - 89%
Kong, 2021	MDD	antidepressants	response: ≥ 50% ↓ HDRS-24	82	40/42	Pearson correlation per sliding window	whole-brain between-ROI FCs	spatio-temporal GCN, GCN, deep-auto encoder, random forest, SVM	10-fold CV	89*	total number: 5; varying: classifiers; accuracies: ca. 50% - 90%
Moreno-Ortega, 2019	MDD	ECT	remission: HDRS-24 ≤ 7	18	9/9	no information	5 specific between- & within-ROI FCs: dorsolateral PFC (p9–46v) - Fundal area of the superior temporal sulcus within MT+ Complex, dorsolateral PFC (p9–46v) - MT+ Complex, dorsolateral PFC (46) - subgenual anterior cingulate cortex, connectivity within the ventral stream visual cortex, connectivity within 10r (part of medial prefrontal cortex)	logistic regression	LOOCV	89	total number: 9; varying: combination of features; accuracies: 72% - 89% (mean: 83%)
Pei, 2020	MDD	SSRI/SNRI	response: ≥ 50% ↓ HDRS-6	98	54/44	Pearson correlation	seed-based whole-brain connectivity of 14 ROIs (all l/r): orbital part of superior frontal gyrus, triangular part inferior frontal gyrus, insula, anterior cingulate gyrus, paracingulate gyrus, posterior cingulate gyrus, hippocampus, amygdala	linear SVM with RFE	LOOCV	81*	total number: 2; varying: subset vs. whole-brain analysis; accuracies: 81%
Schultz, 2018	MDD	SSRI/Alpha2-receptor-antagonists/AAP/ CBT	response: ≥ 50% ↓ BDI	21	7/14	Pearson correlation	between-ROI FCs between 13 ROIs: subgenual anterior cingulate cortex (l/r), amygdala (l/r), intraparietal sulcus (l/r), dorsolateral PFC (l/r), anterior insula (l/r), dorsal anterior cingulate cortex, medial PFC, precuneus	polynomial kernel SVM	LOOCV	72**	total number: 13; varying: features; accuracies: 44% - 89%
Sun, 2020	MDD & BPD	ECT	remission: HDRS-17 <7; response: > 50% ↓ HDRS-17;	122	47/75; 71/51	Pearson correlation	whole-brain between-ROI FCs	(multiple) linear regression, applying	LOOCV, 10-fold-CV	67*	total number: 9 varying: binary outcome, validation technique and

(*continued on next page*)

Table 1 (*continued*)

Study	Primary disorder	Treatment	Definition treatment outcome	N	Responders/ nonresponders	Estimating FCs	Input features	Algorithm(s) of the final classifier (s)	Validation method	Best Acc	Information on models tested
								binarization afterwards			features; accuracies: 58% - 75% (mean: 67%)
Tian, 2020	MDD	SSRI	response: ≥ 50% ↓ HDRS-17 after 8 weeks; nonresponse: less than 20% ↓ after 2 weeks OR less than 50% ↓ after 8 weeks	106	56/50	Pearson correlation per sliding window	node flexibilities per ROI	linear SVM	LOOCV, leave-one-site-out	68**	total number: 4; varying: validation technique; accuracies: 69% - 79% (mean: 73%)
van Waarde, 2015	MDD	ECT	remission: MADRS ⩽10	45	25/20	Dual regression	subject-specific spatial maps	linear SVM	LOOCV	85	total number: 25; varying: features; 2 of 25 models got significant
Wu, 2022	MDD	SSRI	remission: HDRS-17 scores ≤ 7	67	28/39	Pearson correlation	between-ROI FCs between 36 emotion regulation regions of 4 networks: network 1: medial superior frontal gyrus (l, BA 8), middle frontal gyrus (r, BA 8), inferior parietal lobule (l/r, BA 40), medial PFC (l, BA 10), middle frontal gyrus (l, BA 6), middle frontal gyrus (r, BA 11), insula (r), cingulate gyrus (r, BA 23), precuneus (r); network 2: inferior frontal gyrus (l/r, BA 47), superior frontal gyrus (l, BA 6), superior temporal gyrus (l, BA 39), middle temporal gyrus (l, no BA), middle frontal gyrus (l, BA 6), superior frontal gyrus (l, BA 9), caudate (l), tuber (r); network 3: amygdala (l/r), fusiform gyrus (l/r, BA 37), thalamus (r), parahippocampal gyrus (l), medial PFC (bilateral, BA 10), inferior occipital gyrus (l, BA 19); network 4: postcentral gyrus (l/r, BA 2), insula (l, BA 13), superior parietal lobule (l, BA 7), cuneus (l, BA 18), middle occipital gyrus (l, BA 19), thalamus (r), precuneus (r, BA 19), posterior cingulate (r, BA 30)	linear SVM	LOOCV	81*	no other models tested
Zhutovsky, 2019	PTSD	CBT/EMDR	response: ≥ 30% ↓ CAPS	44	24/20	Dual regression	subject-specific spatial maps	Gaussian process classifier	10 × 10-fold CV	81	total number: 48; varying: features; 1 of 48 models got significant
Zhutovsky, 2021	(partial) PTSD	CBT/EMDR	response: ≥ 30% ↓ CAPS-CA	40	21/19	Gig ICA, Pearson correlation, partial correlation	subject-specific spatial maps, connectivity between independent components	linear SVM	50 ×5-fold CV	76	total number: 50; varying: features and types of features (within- and between-network connectivity); 1 of 50 models got significant

Only the balanced accuracy (Acc) of the best model of each study is reported. The asterisk denotes studies for which balanced accuracy was calculated from sensitivity and specificity. The double asterisk denotes studies for which a proxy of balanced accuracy was used. Abbreviations: N = Sample size, FC = functional connectivity, Acc = Accuracy, MDD = major depressive disorder, BPD = bipolar disorder, PTSD = post-traumatic stress

disorder, rTMS = repetitive transcranial magnet stimulation, SSRI = selective serotonin reuptake inhibitor, SNRI = serotonin noradrenaline reuptake inhibitor, ECT = electroconvulsive therapy, AAP = atypical antipsychotics, CBT = cognitive behavioral therapy, EMDR = eye movement desensitization and reprocessing, HDRS = Hamilton Depression Rating Scale, MADRS = Montgomery–Åsberg Depression Rating Scale, CAPS = Clinician Administered PTSD scale, CAPS-CA = Clinician-Administered PTSD Scale for Children and Adolescents, Gig ICA = Group-informed component analysis, ROI = Region of interest, ACC = anterior cingulate cortex, PFC = prefrontal cortex, SVM = support vector machine, GCN = graph convolutional network, LOOCV = leave-one-out cross validation, CV = cross validation.

feature generation. Both were part of our research question. The approaches each study took and their categorization can be seen in table S3.

Around one third of studies (n = 5) ignored the theoretically large amount of initially available connectivities by focusing a priori on specific brain areas and/or connectivities selected according to prior literature and theoretical assumptions. Other studies explored whole-brain functional connectivity but streamlined the number of connectivities for investigation. This was achieved by transitioning from the theoretically available voxel-level to either a brain region level, employing atlas-based parcellations (n = 6), or an independent-component level, utilizing data-driven parcellations (n = 3).

In terms of feature reduction after feature generation, the majority of studies employed feature selection techniques (n = 11). Among them, seven studies utilized so-called filter techniques, which use traditional statistical measures such as correlation coefficients or t-tests to rank features based on their capacity to differentiate between groups. Two other studies employed wrapper techniques, wherein the final classifier is trained in an inner loop to select features based on their importance in the classifier (see Brakowski et al., 2017; Guyon and Elisseeff, 2000; Mwangi et al., 2014 for an overview of feature selection techniques). In a separate set of studies, the number of input features for the final classifier was diminished by distributing the features across multiple models (n = 7). Furthermore, three studies implemented diverse methods of dimensionality reduction post feature generation, including principal component analysis, layering within a convolutional graphical network, or aggregation.

3.6. Risk of bias

All studies were rated as having a high risk of bias. The most common reasons were in the analysis domain, including small sample size, univariate feature selection, and data leakage. A summary of risk of bias is depicted in figure S1, the PROBAST rating for each study is presented in table S4. The most frequent problem was a small sample size, as none of the studies met the PROBAST criterion which requires a number of non-responders that is 10 times larger than the number of candidate features. A further, potentially very severe problem was data leakage, occurring in internal validations, when information from the test set "leaks" into the training set and thus information from the test set is used to train the model. Data leakage highly increases the risk of overestimation as the model can use information which would not be available in a naturalistic setting. Here, data leakage occurred as feature selection (Hopman et al., 2021; Moreno-Ortega et al., 2019) and independent component analysis (van Waarde et al., 2015) were performed on the whole data set. Another reason for high risk of bias were univariate feature selection methods. Univariate feature selection methods include any procedure testing single features for statistically significant relations or group-differences without taking multivariability into account. Univariate feature selection can cause both under- and overestimations of performance accuracies (Jong et al., 2021). Underestimation might result as multivariable patterns with high predictive value in machine learning algorithms being able to handle multivariable data might not be selected (Jong et al., 2021). Overestimation can emerge as univariate selection is more biased to singularities in the data (Jong et al., 2021). Another procedure that was not assessed with the PROBAST rating but can also increase risk of bias is the simultaneous testing of several final models. Most studies (8/13) tested more than one final model, varying feature subsets (e.g., Zhutovsky et al., 2019), classifiers (e.g., Tian et al., 2020), and machine-learning pipelines (e.g., Harris et al., 2022). As most studies reported sufficient performance metrics only for their best model(s), we quantitatively summarized the studies` best models` metrics to assess the predictability of treatment outcome – a common procedure in systematic reviews and/or meta-analyses on machine learning (e.g., Bondi et al., 2023; Y. Lee et al., 2018; Vieira et al., 2022). However, performance metrics of the best one of several final models are

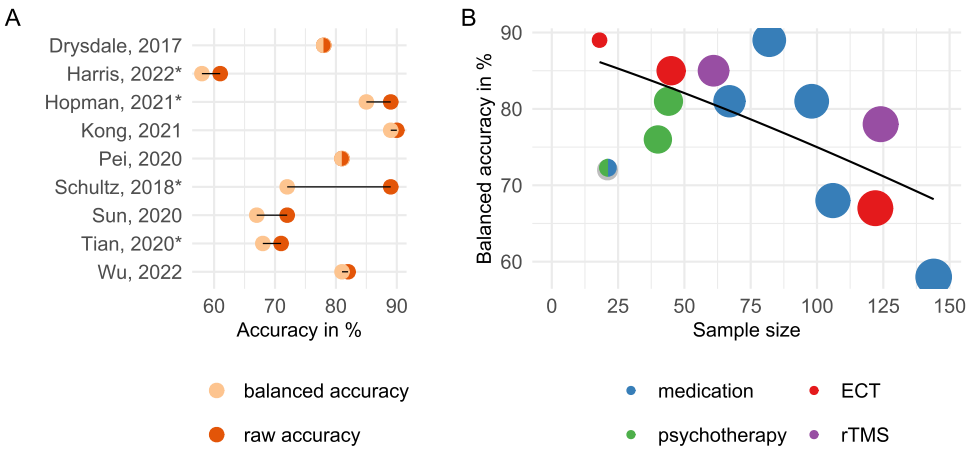

Fig. 2. Difference between raw and balanced accuracies (A) and sample size as moderator in meta-regression (B).A) The difference between balanced and raw accuracy is only depicted for studies that did not report balanced accuracy in the presence of imbalanced classes. The asterix (*) denotes studies where a proxy of balanced accuracy was calculated due to missing information. B) The size of the dots represents the weight of the studies in the meta-analysis. The line is the fitted regression line. Please note that the original meta-regression was performed on the double arcsine transformed proportions. Abbreviations: ECT = electroconvulsive treatment, rTMS = repetitive transcranial magnetic stimulation.

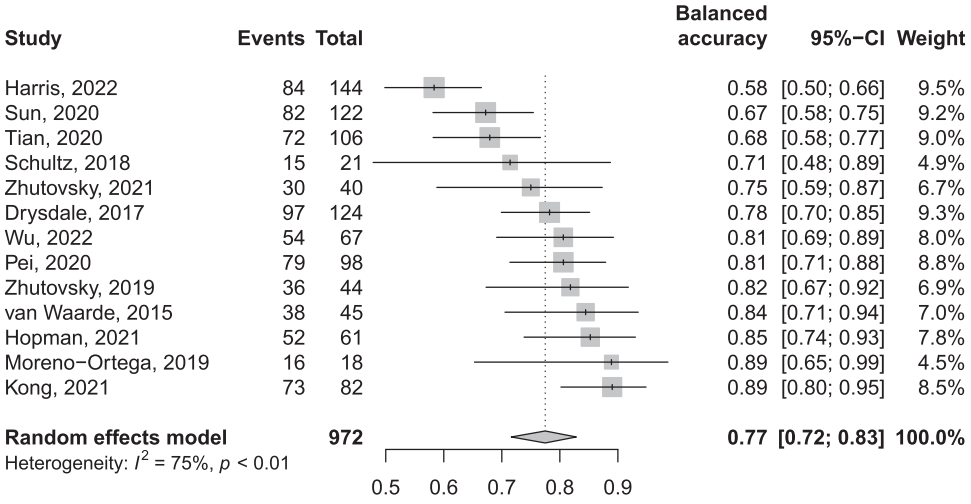

Fig. 3. Forest plot of a random-effect meta-analysis on the balanced accuracy values of the studies` best models.

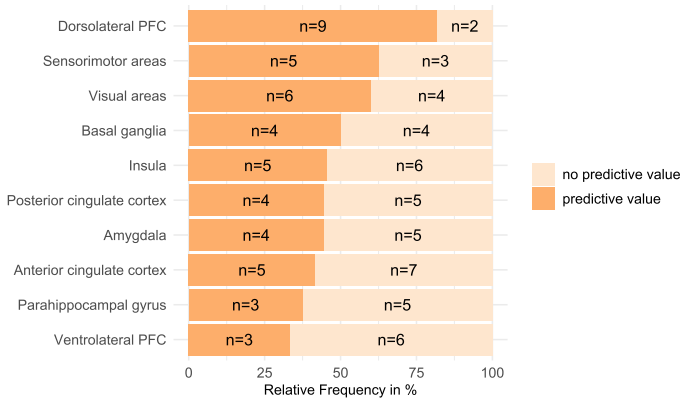

Fig. 4. Absolute and relative frequencies of studies in which a brain area had predictive value. Only brain areas that had predictive value in more than 30% of studies are depicted. The numbers in the bars represent the absolute number of studies in which the brain area had predictive or no predictive value. The brain areas are arranged in descending order following their relative frequency. Abbreviation: PFC = prefrontal cortex.

likely to overestimate the model's performance: In a naturalistic setting, the best model cannot be chosen retrospectively as it should inform the practitioner's further actions before the beginning of treatment.

4. Discussion

The present review and meta-analysis aimed to give an overview of studies using resting-state functional connectivity to predict treatment outcome in internalizing mental disorders. An extensive literature search resulted in 13 studies which predicted outcome to a wide range of treatments, including medication, ECT, psychotherapy, and rTMS and focused mainly on patients with depression. The estimated mean balanced classification accuracy was 77%. A close examination of the connectivities which led to a successful prediction showed that the connectivity of the dorsolateral prefrontal cortex had high predictive value across treatments. The PROBAST rating revealed that all studies suffered from high risk of bias, being especially caused by inappropriate methodological choices.

4.1. Model performance in the light of high risk of bias

The estimated mean balanced classification accuracy of 77% of the

studies' best models reflect that treatment outcome can be predicted better than chance-level. However, the PROBAST rating revealed that all included studies suffered from high risk of bias. The most frequent reasons for high risk of bias were in the analysis domain, including small sample sizes (range: 18 – 144), univariate feature selection methods, and data leakage. Moreover, the simultaneous testing of multiple models represented another source of risk of bias that has not been considered by the PROBAST rating (see a more detailed explanation of these factors in the results section).

Additionally, further exploration of meta-analyses indicated that small sample sizes were associated with elevated classification accuracies, implying a potential overestimation of predictive performance of studies with small-sample sizes. This pattern is consistent with findings in other reviews (e.g., Steele et al., 2018; Vieira et al., 2022). In general, small sample sizes in cross-validation may cause both over- and underestimation of predictive performance on unseen data, as the precision of the model performance, serving as the estimator of model performance on unseen data, decreases with smaller sample sizes (Varoquaux, 2018). The observed association with higher accuracies, both in this study and others, can be attributed to scientific community's tendency to present high classification accuracies (so-called filter effect (Varoquaux, 2018). This interplay is similar to the impact observed in classical statistics with small sample sizes (Button et al., 2013).

Thus, given the high risk of bias across studies and the overrepresentation of small studies likely overestimating the true performance, we consider the estimated mean classification accuracy of 77% as an optimistic upper bound of potential prediction performance rather than a proof of principle. Whether machine learning and functional connectivity are able to predict treatment outcome remains an open question which can only be answered by further studies applying state-of-the-art-machine learning methods lege artis, using larger sample sizes and employing external validation.

4.2. Far from clinical application

Given the high risk of bias previously discussed and particularly the lack of external model validations, our results indicate that models predicting treatment outcome are still far from any clinical application. A similar picture emerges for other mental disorder or neuroimaging variables, with reviews or meta-analyses reporting a wide range of prediction performance across studies, a lack of external validation studies, and, if assessed, a high risk of bias for the included studies (Del Fabro et al., 2023; Vieira et al., 2022; Watts et al., 2022). Thus, our review underscores that the application of machine learning and neuroimaging variables to predict treatment outcome is still far from any clinical application, regardless of mental disorder and neuroimaging modality.

However, even if a model based on resting-state functional connectivity was successfully validated on multiple external datasets, numerous considerations would still precede a clinical application. One of those would be a thorough cost-benefit analysis that considers the specific context and purpose of the application, the model performance, and the monetary costs of collecting the fMRI data. Moreover, aligning with the Research Domain Criteria (RDoC) framework (Cuthbert, 2014), it should be explored whether the functional connectivity patterns which drove the final predictions could be reflected on other units of analysis, such as behavior or physiology, which are more readily assessable. Furthermore, it is important to note that a model which predicts the outcome to a single treatment, as those included here, may not directly contribute to treatment allocation. The direct utility for treatment allocation arises when combining models for different treatments, as demonstrated in approaches like the personalized advantage index (DeRubeis et al., 2014), or when generating models that explicitly recommend one treatment over others. Nevertheless, a model which predicts the outcome to a single treatment holds value by aiding in developing new treatments or add-ons for patients unlikely to respond. Additionally, it could protect patients from undergoing an invasive and high-risk treatment with a low likelihood of success.

4.3. Important brain regions

Connectivities of the DLPFC (here including BA 6, 8, 9, and 46 following the parcellation of Glasser et al., 2016) had high predictive value in the highest number of studies, both in terms of absolute and relative frequency. The DLPFC is part of the central executive network (CEN, also called frontoparietal network; Seeley et al., 2007) that supports decision-making, emotion-regulation, and working memory (Menon, 2011). Connectivity of DLPFC and CEN has been associated with depression (see meta-analysis of Brandl et al., 2022) and has shown treatment-induced changes after ECT and rTMS (see reviews of Brakowski et al., 2017; Porta-Casteràs et al., 2021). Moreover, even though being less consistent, pretreatment connectivity of DLFPC and CEN has been associated with treatment outcome in several studies (see review of Taylor et al., 2021).

Additional support for the hypothesis that the DLPFC might play an important role in the etiology and maintenance of depression comes from lesion-based network mapping, showing that lesions associated with depression can be mapped to a common circuit that is centered in the DLPFC (Padmanabhan et al., 2019). In a similar vein, a recent analysis of task-based fMRI data targeting altered emotional and cognitive processing in depression revealed two robust circuits of altered emotional and cognitive processing which both included the DLPFC (Cash et al., 2023). Interestingly, the abnormal emotion circuit included the left DLPFC, while the abnormal cognition circuit included the right DLPFC, suggesting that a closer look at the DLPFC might be beneficent, both in terms of lateralization and parcellation into subparts (e.g., Cieslik et al., 2013).

Other brain areas with a relative frequency larger 50% were visual and sensorimotor areas. Visual and sensorimotor areas included both lower sensory processing areas, such as the V3 (Drysdale et al., 2017; Tian et al., 2020) and primary sensorimotor cortices (Drysdale et al., 2017; van Waarde et al., 2015; H. Wu et al., 2022) as well as higher sensory processing areas, such as the fusiform faces complex (Sun et al., 2020; H. Wu et al., 2022) and the (pre-)supplementary motor area (Tian et al., 2020; Zhutovsky et al., 2019). Aberrant low- and high-level visual processing and sensorimotor functioning in depression have been reported in several studies, both on a behavioral (e.g., Bennabi et al., 2013; Brakowski et al., 2017; Bubl et al., 2010) and neural level (e.g., Chen et al., 2022; Liu et al., 2022; Ray et al., 2021; Zeng et al., 2012). Moreover, a recent study describing the brain's functional connectivity profile in terms of a principal functional similarity gradient showed that gradient differences between patients with depression and healthy controls were mainly rooted in areas of the visual, sensorimotor and default-mode network (Xia et al., 2022). However, in most recent meta-analyses, alterations in visual or sensorimotor neural processing in depression did not reach significance (e.g., meta-analyses of Brandl et al., 2022; Gray et al., 2020). Additionally, studies investigating and/or reporting associations between pretreatment neural processing in visual or sensorimotor areas and treatment outcome have been rare. Although Dichter et al. (2015) reported pretreatment connectivity differences in visual recognition circuits between responders and non-responders as one of their key findings, this pattern only emerged in 4 of the 21 studies reviewed. Thus, together with current literature, the relatively high predictive value of visual and sensorimotor areas suggests that functional connectivity of these areas plays a role in psychopathology and treatment outcome of depression but might be more difficult to detect or might only be relevant for a subgroup of patients.

The basal ganglia, which include subcortical nuclei like caudate, putamen (striatum) and globus pallidus, showed predictive value in half of the possible studies. Interestingly, three out of the four studies in which the basal ganglia had no predictive value used independent component analysis (ICA), suggesting that ICA might be less suitable to

detect connectivities of the basal ganglia. Indeed, even though a basal ganglia network, comprising basal ganglia components and the thalamus, can be detected in ICA (Robinson et al., 2009), this is often not the case, possibly because of the low proportion of variance it explains (Robinson et al., 2009). Thus, considering the basal ganglias` involvement in cognitive, emotional, and reward processing (e.g., Chakrabarty et al., 2016; Pierce and Péron, 2020), their association with anhedonia (e.g., Borsini et al., 2020; Brandl et al., 2022; Gray et al., 2020) and the target role of ventral striatum and nucleus accumbens in deep brain stimulation (Drobisz and Damborská, 2019; Y. Wu et al., 2021), we suggest to explore the predictive value of basal ganglia functional connectivity with alternative methods.

It is noteworthy that the functional connectivity of other areas commonly involved in the etiology of depression, such as amygdala, insula and anterior cingulate cortex, showed no high predictive value in our analysis. This underscores that neurological correlates of mental disorders may not inherently predict treatment outcome. Other factors, such as plasticity and compensatory neurological mechanisms, could play a more pivotal role. Additionally, capacities like emotion regulation, with distinct correlates from the disorder, may significantly contribute to predicting treatment outcome.

Furthermore, it is important to note that our examination was limited to brain areas whose connectivity demonstrated high predictive value across treatments. The investigation of treatment-specific connectivities with high predictive value was not feasible due to the limited number of studies and high treatment heterogeneity. As a result, our analysis focused solely on identifying brain regions whose connectivities might predict treatment outcome regardless of treatment type. Following the concept outlined by Simon and Perlis (2010), these connectivities could be characterized as general predictors of prognosis or general predictors of treatment response. According to Simon and Perlis (2010), distinguishing between these two groups requires studies predicting response to placebo treatment. If predictors also showed high predictive value in placebo studies, they would be considered general predictors of prognosis; if not, they could be seen as general predictors of treatment response. However, as none of the studies in our analysis used a placebo, we were unable to make this distinction. Nevertheless, irrespective of the accurate characterization, which is less crucial from a machine learning perspective, functional connectivities of the DLPFC, visual, and sensorimotor areas appear to contribute significantly to the correct prediction of machine learning models and should therefore be considered in future models.

4.4. Methodological approaches and recommendation

We summarized current methodological approaches to assist future researchers in making informed decisions regarding two questions: 1. Which approaches are taken to draw inferences about the predictive value of specific features? (RQ 2.1), 2. Which approaches are taken to reduce the large amount of theoretically initially available functional connectivities to a small set of features to be used in the final classifier (s)? (RQ 3)

Regarding the first question (RQ 2.1), approaches taken to draw inferences about the featureś predictive value, we identified three distinct groups of methodologies that were mostly employed exclusively: model comparison, selection frequency in feature selection, and feature importance in final classifier. As these approaches evaluate the predictive value of features on distinct levels, we refrain from favoring one over the other and view them as complementary. For instance, a high selection frequency in feature selection indicates that a feature was included in the final model in most iterations but does not necessarily imply that this feature also drove the final prediction. We recommend, therefore, deviating from common practice by assessing predictive value on multiple levels when suitable. We advise consistently evaluating feature importance in the final classifier and to assess selection frequency when employing feature selection, even when the primary goal is to compare the predictive value of different features by comparing distinct models (for a review of different measures of feature importance see Mi et al., 2020). This approach ensures a comprehensive understanding of predictive value and helps to detect potential errors or unexpected model behavior.

Concerning the second methodological question (RQ 3), approaches taken to reduce the large amount of theoretically initially available functional connectivities, we observed a common procedure, that accommodated a diverse array of approaches. First, all studies performed some kind of initial reduction before feature generation by either selecting specific brain areas and/or connectivities or parcellating the whole-brain data in a data- or atlas-based manner. Second, after generating functional-connectivity based features (e.g., functional connectivities themselves, node flexibilities, or subject-specific spatial maps), the majority of studies further diminished the number of input features by employing feature selection techniques (mainly filter techniques) and/or distributing features to multiple models. The specific approaches varied among all studies, even when calculating the same type of features. This methodological heterogeneity underscores the lack of standards in dealing with the large amount of theoretically available functional connectivities making the comparison of study findings more challenging. Therefore, further studies are needed to systematically explore the effects of various methodological choices across different data sets in order to provide general recommendations.

Regarding approaches that involve an initial reduction before feature generation, we refrain from providing explicit recommendations. Both methods, namely a priori selection of functional connectivities and whole-brain analysis with a parcellation technique, have their respective advantages and drawbacks. While the number of features generated after a priori selection may be more manageable for machine learning within the current scope, this approach may only be valid if the selection is thoroughly justified, which is often not the case.

In contrast, regarding approaches that serve feature reduction after feature generation, we would like to highlight several shortcomings that should be addressed in future studies. First, the most commonly employed approach, feature selection via filter techniques, increases the risk of bias due to the inherent univariance of filter techniques. This risk of bias emerges as these techniques are unable to select multivariate patterns with high predictive value (Mwangi et al., 2014). Hence, we recommend employing more sophisticated feature selection techniques such as wrapper and embedded methods, being more suitable for multivariate data (Mwangi et al., 2014). Second, the other frequently utilized approach, allocating features to different models, also amplifies the risk of bias in its typical implementation. Most studies tested several models in parallel and reported the prediction accuracy of the best model as estimate of prediction performance. As pointed out in the results section, this procedure might induce bias, as in a naturalistic setting, the best model cannot be chosen retrospectively; it should inform the practitioneŕs further actions before the beginning of treatment. To reduce the risk of bias without employing an additional external validation, we recommend to train one final second-level model on the predictions of several first-level models, as applied by Pei et al. (2020).

5. Limitations

Our review has several limitations. First, even though commonly applied (e.g., Y. Lee et al., 2018; Vieira et al., 2022), the suitability of using meta-analysis for proportions in synthesizing cross-validation accuracies is not conclusively established. Potential issues emerge in terms of estimating the error variance (the square of standard measurement error) which is typically used to weight the studies' results in the meta-analysis. In general, there might be no unbiased estimator of error variance for classification accuracy in cross-validation (for an overview of methods to estimate the error variance in cross-validation see Bates et al., 2023). Moreover, the use of an error variance estimator originally

developed for proportions overlooks the impact of certain cross-validation types, like leave-one-out, on increasing error variance (Varoquaux, 2018). Additionally, assumptions underlying the estimation of error variance for proportions, such as each subject having an equal chance of being a case, do not hold in cross-validation, where the probability of being a case (= being correctly classified) varies for each training-test split. Despite these limitations, we maintain that this meta-analysis remains valuable in providing a numerical estimate of the current predictive capability for treatment outcome. Future research should dedicate attention to addressing this issue and developing guidelines for conducting meta-analyses of classification accuracies based on cross-validation.

Second, all studies included had a high risk of bias. Therefore, the results presented here should be interpreted with caution. Please note however that this caveat applies to most systematic reviews in precision healthcare as the PROBAST rating has revealed high risk of bias of predictive modelling in healthcare, notably due to methodological weaknesses in the analysis domain (Jong et al., 2021; Meehan et al., 2022). Third, even though we initially intended to give an overview about the current state of treatment outcome prediction in a wide spectrum of internalizing mental disorders, including also anxiety disorders and obsessive-compulsive disorders, we eventually only included studies on depression and post-traumatic stress disorder. This was not due to a lack of studies predicting treatment outcome in anxiety disorders and obsessive-compulsive disorders per se, but due to a lack of studies fulfilling our inclusion criteria, such as using a machine learning approach (Chen et al., 2022; Göttlich et al., 2015) or applying a model being only based on resting-state functional connectivity (Reggente et al., 2018; Whitfield-Gabrieli et al., 2016). The sample of included studies is thus not fully representative of the entire spectrum of internalizing mental disorders, and it is not clear to which extent our results are also valid for other, initially targeted internalized disorders. Fourth, we summarized important brain regions based on study-specific labels instead of using more sophisticated approaches such as coordinate-based meta-analysis, because a substantial proportion (4 out of 12) of the studies included lacked adequate coordinate information, mainly due to using methods that typically do not provide such information. Additionally, it is important to note that the initial screening was conducted by a single individual, which may not adhere to gold-standard practices.

6. Conclusion and further directions

The objective of this review was to provide a comprehensive overview of studies utilizing resting-state functional connectivity to predict treatment outcomes in internalizing mental disorders across a spectrum of treatments, encompassing psychotherapy, pharmacotherapy, rTMS, and ECT. Our meta-analysis indicated that treatment outcome can be predicted based on resting-state functional connectivity, with a mean estimated balanced accuracy of 77% (95% CI: [72%- 83%]). However, aiming to give a realistic estimate of the potential of machine learning and resting-state functional connectivity, we underscored the need to interpret these values cautiously, considering them more as an optimistic upper limit of potential prediction performance. This caution stems from the influence of small sample sizes systematically biasing the results and a notable risk of bias, as evaluated through PROBAST. A closer look at connectivities which drove a successful prediction highlighted the important role of the dorsolateral prefrontal cortex and raised awareness of two other, previously rather neglected groups of brain areas: visual and sensorimotor areas. In future studies conducting an a priori feature selection for predicting treatment outcomes, it is advisable to contemplate the inclusion of functional connectivities from these specific areas. Moreover, summarizing current methodological practices and employing PROBAST, we have identified several methodological choices that should be considered in future studies. These include the use of larger sample sizes, potentially through collaborative efforts, evaluating predictive value on several levels, opting for multivariable instead of univariate feature selection techniques, and generating one final 2nd-level model when comparing models based on different feature sets.

Besides these methodological choices, current developments might further leverage the predictive ability of resting-state functional connectivity. First, state-of the-art methods to estimate functional connectivities might lead to better and more robust predictions (see review here; Colclough et al., 2018), mitigating the problems of the typically used Pearson correlations which suffer from a low signal-to-noise-ratio (Pervaiz et al., 2020) and lack a distinction between direct and indirect connectivities (Smith et al., 2011). Second, another promising avenue might be measures which summarize a regions` or networks` connectivity in an informative manner, such as graph metrics (Rubinov and Sporns, 2010), circuit scores (e.g., Goldstein-Piekarski et al., 2022), and functional-similarity gradients (Haak et al., 2018). Lastly, another fruitful development could involve characterizing single-subject connectivity by framing them as deviations from those observed in healthy controls. Both straightforward methods, such as quantifying measures as z-deviations (Goldstein-Piekarski et al., 2022), and more complex methods, such as normative modelling (Marquand et al., 2019), might be beneficent. The approaches have the potential to better capture inter-subject heterogeneity of functional connectivity and to reduce the impact of noise.

Data availability

The R-code used to perform the analyses and to create the presented plots and tables as well as all our extracted data are available in our OSF-repository (https://osf.io/y69ke/).

Acknowledgements

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - FOR5187 (project number 442075332).

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.neubiorev.2024.105640.

References

Andrews, G., 2018. Internalizing disorders: The whole is greater than the sum of the parts. World Psychiatry.: Off. J. World Psychiatr. Assoc. (WPA) *17* (3), 302–303. https://doi.org/10.1002/wps.20564.

Balduzzi, S., Rücker, G., Schwarzer, G., 2019. How to perform a meta-analysis with R: A practical tutorial. Evid. -Based Ment. Health *22* (4), 153–160. https://doi.org/10.1136/ebmental-2019-300117.

Barendregt, J.J., Doi, S.A., Lee, Y.Y., Norman, R.E., Vos, T., 2013. Meta-analysis of prevalence. J. Epidemiol. Community Health *67* (11), 974–978. https://doi.org/10.1136/jech-2013-203104.

Bates, S., Hastie, T., Tibshirani, R., 2023. Cross-Validation: What Does It Estimate and How Well Does It Do It? J. Am. Stat. Assoc. 1–12. https://doi.org/10.1080/01621459.2023.2197686.

Bennabi, D., Vandel, P., Papaxanthis, C., Pozzo, T., Haffen, E., 2013. Psychomotor retardation in depression: A systematic review of diagnostic, pathophysiologic, and therapeutic implications. BioMed. Res. Int. 2013 https://doi.org/10.1155/2013/158746. Article 158746.

Bondi, E., Maggioni, E., Brambilla, P., Delvecchio, G., 2023. A systematic review on the potential use of machine learning to classify major depressive disorder from healthy controls using resting state fMRI measures. Neurosci. Biobehav. Rev. 144 https://doi.org/10.1016/j.neubiorev.2022.104972. Article 104972.

Borsini, A., Wallis, A.S.J., Zunszain, P., Pariante, C.M., Kempton, M.J., 2020. Characterizing anhedonia: A systematic review of neuroimaging across the subtypes of reward processing deficits in depression. Cogn., Affect., Behav. Neurosci. *20* (4), 816–841. https://doi.org/10.3758/s13415-020-00804-6.

Brakowski, J., Spinelli, S., Dörig, N., Bosch, O.G., Manoliu, A., Holtforth, M.G., Seifritz, E., 2017. Resting state brain network function in major depression - Depression symptomatology, antidepressant treatment effects, future research. J. Psychiatr. Res. *92*, 147–159. https://doi.org/10.1016/j.jpsychires.2017.04.007.

Brandl, F., Weise, B., Mulej Bratec, S., Jassim, N., Hoffmann Ayala, D., Bertram, T., Ploner, M., Sorg, C., 2022. Common and specific large-scale brain changes in major depressive disorder, anxiety disorders, and chronic pain: A transdiagnostic multimodal meta-analysis of structural and functional MRI studies. Neuropsychopharmacology *47* (5), 1071–1080. https://doi.org/10.1038/s41386-022-01271-y.

Brodersen, K.H., Ong, C.S., Stephan, K.E., Buhmann, J.M., 2010. The Balanced Accuracy and Its Posterior Distribution. *2010 20th International Conference on Pattern Recognition.* IEEE. https://doi.org/10.1109/icpr.2010.764.

Bubl, E., Kern, E., Ebert, D., Bach, M., van Tebartz Elst, L., 2010. Seeing gray when feeling blue? Depression can be measured in the eye of the diseased. Biol. Psychiatry *68* (2), 205–208. https://doi.org/10.1016/j.biopsych.2010.02.009.

Carpenter, J.K., Andrews, L.A., Witcraft, S.M., Powers, M.B., Smits, J.A.J., Hofmann, S. G., 2018. Cognitive behavioral therapy for anxiety and related disorders: A meta-analysis of randomized placebo-controlled trials. Depress Anxiety 35 (6), 502–514. https://doi.org/10.1002/da.22728.

Cash, R.F.H., Müller, V.I., Fitzgerald, P.B., Eickhoff, S.B., Zalesky, A., 2023. Altered brain activity in unipolar depression unveiled using connectomics. Nat. Ment. Health *1* (3), 174–185. https://doi.org/10.1038/s44220-023-00038-8.

Chakrabarty, T., Ogrodniczuk, J., Hadjipavlou, G., 2016. Predictive neuroimaging markers of psychotherapy response: A systematic review. Harv. Rev. Psychiatry *24* (6), 396–405. https://doi.org/10.1097/HRP.0000000000000132.

Chen, X., Wang, Z., Lv, Q., Lv, Q., van Wingen, G., Fridgeirsson, E.A., Denys, D., Voon, V., Wang, Z., 2022. Common and differential connectivity profiles of deep brain stimulation and capsulotomy in refractory obsessive-compulsive disorder. Mol. Psychiatry 27, 1020–1030. https://doi.org/10.1038/s41380-021-01358-w.

Cieslik, E.C., Zilles, K., Caspers, S., Roski, C., Kellermann, T.S., Jakobs, O., Langner, R., Laird, A.R., Fox, P.T., Eickhoff, S.B., 2013. Is there "one" DLPFC in cognitive action control? Evidence for heterogeneity from co-activation-based parcellation. Cereb. Cortex *23* (11), 2677–2689. https://doi.org/10.1093/cercor/bhs256.

Cohen, S.E., Zantvoord, J.B., Wezenberg, B.N., Bockting, C.L.H., van Wingen, G.A., 2021. Magnetic resonance imaging for individual prediction of treatment response in major depressive disorder: A systematic review and meta-analysis. Transl. Psychiatry 11 (1). https://doi.org/10.1038/s41398-021-01286-x. Article 168.

Colclough, G.L., Woolrich, M.W., Harrison, S.J., Rojas López, P.A., Valdes-Sosa, P.A., Smith, S.M., 2018. Multi-subject hierarchical inverse covariance modelling improves estimation of functional brain networks. NeuroImage *178*, 370–384. https://doi.org/10.1016/j.neuroimage.2018.04.077.

Cuijpers, P., Sijbrandij, M., Koole, S.L., Andersson, G., Beekman, A.T., Reynolds, C.F., 2013. The efficacy of psychotherapy and pharmacotherapy in treating depressive and anxiety disorders: A meta-analysis of direct comparisons. World Psychiatry *12* (2), 137–148. https://doi.org/10.1002/wps.20038.

Cuthbert, B.N., 2014. The RDoC framework: Facilitating transition from ICD/DSM to dimensional approaches that integrate neuroscience and psychopathology. World Psychiatry.: Off. J. World Psychiatr. Assoc. (WPA) *13* (1), 28–35. https://doi.org/10.1002/wps.20087.

Dalhuisen, I., van Bronswijk, S., Bors, J., Smit, F., Spijker, J., Tendolkar, I., Ruhé, H.G., van Eijndhoven, P., 2022. The association between sample and treatment characteristics and the efficacy of repetitive transcranial magnetic stimulation in depression: A meta-analysis and meta-regression of sham-controlled trials. Neurosci. Biobehav. Rev. 141 https://doi.org/10.1016/j.neubiorev.2022.104848. Article 104848.

Del Fabro, L., Bondi, E., Serio, F., Maggioni, E., D'Agostino, A., Brambilla, P., 2023. Machine learning methods to predict outcomes of pharmacological treatment in psychosis. Transl. Psychiatry 13 (1). https://doi.org/10.1038/s41398-023-02371-z. Article 75.

DeRubeis, R.J., 2019. The history, current status, and possible future of precision mental health. Behav. Res. Ther. 123 https://doi.org/10.1016/j.brat.2019.103506. Article 103506.

DeRubeis, R.J., Cohen, Z.D., Forand, N.R., Fournier, J.C., Gelfand, L.A., Lorenzo-Luaces, L., 2014. The Personalized Advantage Index: Translating research on prediction into individualized treatment recommendations. A demonstration. PLoS ONE 9 (1). https://doi.org/10.1371/journal.pone.0083875. Article e83875.

Dichter, G.S., Gibbs, D., Smoski, M.J., 2015. A systematic review of relations between resting-state functional-MRI and treatment response in major depressive disorder. J. Affect. Disord. *172*, 8–17. https://doi.org/10.1016/j.jad.2014.09.028.

Drobisz, D., Damborská, A., 2019. Deep brain stimulation targets for treating depression. Behav. Brain Res. *359*, 266–273. https://doi.org/10.1016/j.bbr.2018.11.004.

Drysdale, A.T., Grosenick, L., Downar, J., Dunlop, K., Mansouri, F., Meng, Y., Fetcho, R. N., Zebley, B., Oathes, D.J., Etkin, A., Schatzberg, A.F., Sudheimer, K., Keller, J., Mayberg, H.S., Gunning, F.M., Alexopoulos, G.S., Fox, M.D., Pascual-Leone, A., Voss, H.U., Liston, C., 2017. Resting-state connectivity biomarkers define neurophysiological subtypes of depression. Nat. Med. *23* (1), 28–38. https://doi.org/10.1038/nm.4246.

Fitzgerald, P.B., 2020. An update on the clinical use of repetitive transcranial magnetic stimulation in the treatment of depression. J. Affect. Disord. *276*, 90–103. https://doi.org/10.1016/j.jad.2020.06.067.

Fonseka, T.M., MacQueen, G.M., Kennedy, S.H., 2018. Neuroimaging biomarkers as predictors of treatment outcome in Major Depressive Disorder. J. Affect. Disord. *233*, 21–35. https://doi.org/10.1016/j.jad.2017.10.049.

GBD 2019 Mental Disorders Collaborators, 2022. Global, regional, and national burden of 12 mental disorders in 204 countries and territories, 1990-2019: A systematic analysis for the Global Burden of Disease Study 2019. Lancet Psychiatry *9* (2), 137–150. https://doi.org/10.1016/S2215-0366(21)00395-3.

Glasser, M.F., Coalson, T.S., Robinson, E.C., Hacker, C.D., Harwell, J., Yacoub, E., Ugurbil, K., Andersson, J., Beckmann, C.F., Jenkinson, M., Smith, S.M., van Essen, D. C., 2016. A multi-modal parcellation of human cerebral cortex. Nature *536*, 171–178. https://doi.org/10.1038/nature18933.

Goldstein-Piekarski, A.N., Ball, T.M., Samara, Z., Staveland, B.R., Keller, A.S., Fleming, S. L., Grisanzio, K.A., Holt-Gosselin, B., Stetz, P., Ma, J., Williams, L.M., 2022. Mapping neural circuit biotypes to symptoms and behavioral dimensions of depression and anxiety. Biol. Psychiatry 91 (6), 561–571. https://doi.org/10.1016/j.biopsych.2021.06.024.

Göttlich, M., Krämer, U.M., Kordon, A., Hohagen, F., Zurowski, B., 2015. Resting-state connectivity of the amygdala predicts response to cognitive behavioral therapy in obsessive compulsive disorder. Biol. Psychol. *111*, 100–109. https://doi.org/10.1016/j.biopsycho.2015.09.004.

Gray, J.P., Müller, V.I., Eickhoff, S.B., Fox, P.T., 2020. Multimodal abnormalities of brain structure and function in major depressive disorder: A meta-analysis of neuroimaging studies. Am. J. Psychiatry *177* (5), 422–434. https://doi.org/10.1176/appi.ajp.2019.19050560.

Guyon, I., Elisseeff, A., 2000. 10.1162/153244303322753616. J. Mach. Learn. Res. *3*, 1157–1182. https://doi.org/10.1162/153244303322753616.

Haak, K.V., Marquand, A.F., Beckmann, C.F., 2018. Connectopic mapping with resting-state fMRI. NeuroImage *170*, 83–94. https://doi.org/10.1016/j.neuroimage.2017.06.075.

Harris, J.K., Hassel, S., Davis, A.D., Zamyadi, M., Arnott, S.R., Milev, R., Lam, R.W., Frey, B.N., Hall, G.B., Müller, D.J., Rotzinger, S., Kennedy, S.H., Strother, S.C., MacQueen, G.M., Greiner, R., 2022. Predicting escitalopram treatment response from pre-treatment and early response resting state fMRI in a multi-site sample: A CAN-BIND-1 report. NeuroImage: Clin. 35 https://doi.org/10.1016/j.nicl.2022.103120. Article 103120.

Hettema, J.M., Neale, M.C., Myers, J.M., Prescott, C.A., Kendler, K.S., 2006. A population-based twin study of the relationship between neuroticism and internalizing disorders. Am. J. Psychiatry *163* (5), 857–864. https://doi.org/10.1176/ajp.2006.163.5.857.

Higgins, J.P.T., Thompson, S.G., 2002. Quantifying heterogeneity in a meta-analysis. Stat. Med. *21* (11), 1539–1558. https://doi.org/10.1002/sim.1186.

Hopman, H.J., Chan, S.M.S., Chu, W.C.W., Lu, H., Tse, C.-Y., Chau, S.W.H., Lam, L.C.W., Mak, A.D.P., Neggers, S.F.W., 2021. Personalized prediction of transcranial magnetic stimulation clinical response in patients with treatment-refractory depression using neuroimaging biomarkers and machine learning. J. Affect. Disord. *290*, 261–271. https://doi.org/10.1016/j.jad.2021.04.081.

Huang, C.-C., Rolls, E.T., Feng, J., Lin, C.-P., 2022. An extended Human Connectome Project multimodal parcellation atlas of the human cortex and subcortical areas. Brain Struct. Funct. *227* (3), 763–778. https://doi.org/10.1007/s00429-021-02421-6.

Ioannidis, J.P.A., Trikalinos, T.A., 2007. The appropriateness of asymmetry tests for publication bias in meta-analyses: A large survey. CMAJ *176* (8), 1091–1096. https://doi.org/10.1503/cmaj.060410.

Jong, Y., de, Ramspek, C.L., Zoccali, C., Jager, K.J., Dekker, F.W., van Diepen, M., 2021. Appraising prediction research: A guide and meta-review on bias and applicability assessment using the Prediction model Risk Of Bias ASsessment Tool (PROBAST). Nephrology *26* (12), 939–947. https://doi.org/10.1111/nep.13913.

Kang, S.-G., Cho, S.-E., 2020. Neuroimaging Biomarkers for Predicting Treatment Response and Recurrence of Major Depressive Disorder. Int. J. Mol. Sci. 21 (6) https://doi.org/10.3390/ijms21062148. Article 2148.

Karvelis, P., Charlton, C.E., Allohverdi, S.G., Bedford, P., Hauke, D.J., Diaconescu, A.O., 2022. Computational approaches to treatment response prediction in major depression using brain activity and behavioral data: A systematic review. Netw. Neurosci. *6* (4), 1066–1103. https://doi.org/10.1162/netn_a_00233.

Kessler, R.C., Ormel, J., Petukhova, M., McLaughlin, K.A., Green, J.G., Russo, L.J., Stein, D.J., Zaslavsky, A.M., Aguilar-Gaxiola, S., Alonso, J., Andrade, L., Benjet, C., Girolamo, G., de, Graaf, R., de, Demyttenaere, K., Fayyad, J., Haro, J.M., Hu, C. y, Karam, A., Ustün, T.B., 2011. Development of lifetime comorbidity in the World Health Organization world mental health surveys. Arch. Gen. Psychiatry *68* (1), 90–100. https://doi.org/10.1001/archgenpsychiatry.2010.180.

Khosla, M., Jamison, K., Ngo, G.H., Kuceyeski, A., Sabuncu, M.R., 2019. Machine learning in resting-state fMRI analysis. Magn. Reson. Imaging *64*, 101–121. https://doi.org/10.1016/j.mri.2019.05.031.

Knapp, G., Hartung, J., 2003. Improved tests for a random effects meta-regression with a single covariate. Stat. Med. *22* (17), 2693–2710. https://doi.org/10.1002/sim.1482.

Kong, Y., Gao, S., Yue, Y., Hou, Z., Shu, H., Xie, C., Zhang, Z., Yuan, Y., 2021. Spatio-temporal graph convolutional network for diagnosis and treatment response prediction of major depressive disorder from functional connectivity. Hum. Brain Mapp. 42 (12), 3922–3933. https://doi.org/10.1002/hbm.25529.

Kotov, R., Krueger, R.F., Watson, D., Achenbach, T.M., Althoff, R.R., Bagby, R.M., Brown, T.A., Carpenter, W.T., Caspi, A., Clark, L.A., Eaton, N.R., Forbes, M.K., Forbush, K.T., Goldberg, D., Hasin, D., Hyman, S.E., Ivanova, M.Y., Lynam, D.R., Markon, K., Zimmerman, M., 2017. The hierarchical taxonomy of psychopathology (HiTOP): a dimensional alternative to traditional nosologies. J. Abnorm. Psychol. *126* (4), 454–477. https://doi.org/10.1037/abn0000258.

Lee, J., Chi, S., Lee, M.-S., 2022. Personalized diagnosis and treatment for neuroimaging in depressive disorders. J. Pers. Med. 12 (9) https://doi.org/10.3390/jpm12091403. Article 1403.

Lee, Y., Ragguett, R.-M., Mansur, R.B., Boutilier, J.J., Rosenblat, J.D., Trevizol, A., Brietzke, E., Lin, K., Pan, Z., Subramaniapillai, M., Chan, T.C.Y., Fus, D., Park, C., Musial, N., Zuckerman, H., Chen, V.C.-H., Ho, R., Rong, C., McIntyre, R.S., 2018. Applications of machine learning algorithms to predict therapeutic outcomes in depression: a meta-analysis and systematic review. J. Affect. Disord. *241*, 519–532. https://doi.org/10.1016/j.jad.2018.08.073.

Liu, X., Klugah-Brown, B., Zhang, R., Chen, H., Zhang, J., Becker, B., 2022. Pathological fear, anxiety and negative affect exhibit distinct neurostructural signatures: evidence from psychiatric neuroimaging meta-analysis. Transl. Psychiatry 12 (1). https://doi.org/10.1038/s41398-022-02157-9. Article 405.

Loerinc, A.G., Meuret, A.E., Twohig, M.P., Rosenfield, D., Bluett, E.J., Craske, M.G., 2015. Response rates for CBT for anxiety disorders: need for standardized criteria. Clin. Psychol. Rev. *42*, 72–82. https://doi.org/10.1016/j.cpr.2015.08.004.

Lv, H., Wang, Z., Tong, E., Williams, L.M., Zaharchuk, G., Zeineh, M., Goldstein-Piekarski, A.N., Ball, T.M., Liao, C., Wintermark, M., 2018. Resting-state functional mri: everything that nonexperts have always wanted to know. Am. J. Neuroradiol. 39 (8), 1390–1399. https://doi.org/10.3174/ajnr.A5527.

Mack, S., Jacobi, F., Beesdo-Baum, K., Gerschler, A., Strehle, J., Höfler, M., Busch, M.A., Maske, U., Hapke, U., Gaebel, W., Zielasek, J., Maier, W., Wittchen, H.U., 2015. Functional disability and quality of life decrements in mental disorders: results from the mental health module of the german health interview and examination survey for adults (DEGS1-MH). Eur. Psychiatry *30* (6), 793–800. https://doi.org/10.1016/j.eurpsy.2015.06.003.

Marquand, A.F., Kia, S.M., Zabihi, M., Wolfers, T., Buitelaar, J.K., Beckmann, C.F., 2019. Conceptualizing mental disorders as deviations from normative functioning. Mol. Psychiatry *24*, 1415–1424. https://doi.org/10.1038/s41380-019-0441-1.

Meehan, A.J., Lewis, S.J., Fazel, S., Fusar-Poli, P., Steyerberg, E.W., Stahl, D., Danese, A., 2022. Clinical prediction models in psychiatry: a systematic review of two decades of progress and challenges. Mol. Psychiatry *27*, 2700–2708. https://doi.org/10.1038/s41380-022-01528-4.

Menon, V., 2011. Large-scale brain networks and psychopathology: a unifying triple network model. Trends Cogn. Sci. *15* (10), 483–506. https://doi.org/10.1016/j.tics.2011.08.003.

Mi, J.-X., Li, A.-D., Zhou, L.-F., 2020. Review study of interpretation methods for future interpretable machine learning. IEEE Access *8*, 191969–191985. https://doi.org/10.1109/access.2020.3032756.

Moreno-Ortega, M., Prudic, J., Rowny, S., Patel, G.H., Kangarlu, A., Lee, S., Grinband, J., Palomo, T., Perera, T., Glasser, M.F., Javitt, D.C., 2019. Resting state functional connectivity predictors of treatment response to electroconvulsive therapy in depression. Sci. Rep. 9 (1) https://doi.org/10.1038/s41598-019-41175-4. Article 5071.

Mutz, J., Vipulananthan, V., Carter, B., Hurlemann, R., Fu, C.H.Y., Young, A.H., 2019. Comparative efficacy and acceptability of non-surgical brain stimulation for the acute treatment of major depressive episodes in adults: systematic review and network meta-analysis. BMJ 364. https://doi.org/10.1136/bmj.l1079. Article 1079.

Mwangi, B., Tian, T.S., Soares, J.C., 2014. A review of feature reduction techniques in neuroimaging. Neuroinformatics *12* (2), 229–244. https://doi.org/10.1007/s12021-013-9204-3.

Padmanabhan, J.L., Cooke, D., Joutsa, J., Siddiqi, S.H., Ferguson, M., Darby, R.R., Soussand, L., Horn, A., Kim, N.Y., Voss, J.L., Naidech, A.M., Brodtmann, A., Egorova, N., Gozzi, S., Phan, T.G., Corbetta, M., Grafman, J., Fox, M.D., 2019. A human depression circuit derived from focal brain lesions. Biol. Psychiatry *86* (10), 749–758. https://doi.org/10.1016/j.biopsych.2019.07.023.

Papakostas, G.I., Fava, M., 2009. Does the probability of receiving placebo influence clinical trial outcome? A meta-regression of double-blind, randomized clinical trials in MDD. Eur. Neuropsychopharmacol. *19* (1), 34–40. https://doi.org/10.1016/j.euroneuro.2008.08.009.

Pei, C., Sun, Y., Zhu, J., Wang, X., Zhang, Y., Zhang, S., Yao, Z., Lu, Q., 2020. Ensemble learning for early-response prediction of antidepressant treatment in major depressive disorder. J. Magn. Reson. Imaging 52 (1), 161–171. https://doi.org/10.1002/jmri.27029.

Pervaiz, U., Vidaurre, D., Woolrich, M.W., Smith, S.M., 2020. Optimising network modelling methods for fMRI. NeuroImage 211. https://doi.org/10.1016/j.neuroimage.2020.116604. Article 116604.

Pierce, J.E., Péron, J., 2020. The basal ganglia and the cerebellum in human emotion. Soc. Cogn. Affect. Neurosci. *15* (5), 599–613. https://doi.org/10.1093/scan/nsaa076.

Porta-Casteràs, D., Cano, M., Camprodon, J.A., Loo, C., Palao, D., Soriano-Mas, C., Cardoner, N., 2021. A multimetric systematic review of fMRI findings in patients with MDD receiving ECT. Prog. Neuro-Psychopharmacol. Biol. Psychiatry 108. https://doi.org/10.1016/j.pnpbp.2020.110178. Article 110178.

R Core Team. (2020). *R: A language and environment for statistical computing* (Version 4.3.0) [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/.

Ray, D., Bezmaternykh, D., Mel'nikov, M., Friston, K.J., Das, M., 2021. Altered effective connectivity in sensorimotor cortices is a signature of severity and clinical course in depression. Proc. Natl. Acad. Sci. USA 118 (40). https://doi.org/10.1073/pnas.2105730118. Article e2105730118.

Reggente, N., Moody, T.D., Morfini, F., Sheen, C., Rissman, J., O'Neill, J., Feusner, J.D., 2018. Multivariate resting-state functional connectivity predicts response to cognitive behavioral therapy in obsessive-compulsive disorder. Proc. Natl. Acad. Sci. USA *115* (9), 2222–2227. https://doi.org/10.1073/pnas.1716686115.

Robinson, S., Basso, G., Soldati, N., Sailer, U., Jovicich, J., Bruzzone, L., Kryspin-Exner, I., Bauer, H., Moser, E., 2009. A resting state network in the motor control circuit of the basal ganglia. BMC Neurosci. 10 https://doi.org/10.1186/1471-2202-10-137. Article 137.

Rubinov, M., Sporns, O., 2010. Complex network measures of brain connectivity: uses and interpretations. NeuroImage *52* (3), 1059–1069. https://doi.org/10.1016/j.neuroimage.2009.10.003.

Schultz, J., Becker, B., Preckel, K., Seifert, M., Mielacher, C., Conrad, R., Kleiman, A., Maier, W., Kendrick, K.M., Hurlemann, R., 2018. Improving therapy outcome prediction in major depression using multimodal functional neuroimaging: a pilot study. Pers. Med. Psychiatry *11-12*, 7–15. https://doi.org/10.1016/j.pmip.2018.09.001.

Seeley, W.W., Menon, V., Schatzberg, A.F., Keller, J., Glover, G.H., Kenna, H., Reiss, A.L., Greicius, M.D., 2007. Dissociable intrinsic connectivity networks for salience processing and executive control. J. Neurosci. *27* (9), 2349–2356. https://doi.org/10.1523/JNEUROSCI.5587-06.2007.

Sidey-Gibbons, J.A.M., Sidey-Gibbons, C.J., 2019. Machine learning in medicine: a practical introduction. BMC Med. Res. Methodol. 19 (1) https://doi.org/10.1186/s12874-019-0681-4. Article 64.

Simon, G.E., Perlis, R.H., 2010. Personalized medicine for depression: can we match patients with treatments? Am. J. Psychiatry *167* (12), 1445–1455. https://doi.org/10.1176/appi.ajp.2010.09111680.

Smith, S.M., Miller, K.L., Salimi-Khorshidi, G., Webster, M., Beckmann, C.F., Nichols, T.E., Ramsey, J.D., Woolrich, M.W., 2011. Network modelling methods for FMRI. NeuroImage *54* (2), 875–891. https://doi.org/10.1016/j.neuroimage.2010.08.063.

Steele, V.R., Maurer, J.M., Arbabshirani, M.R., Claus, E.D., Fink, B.C., Rao, V., Calhoun, V.D., Kiehl, K.A., 2018. Machine learning of functional magnetic resonance imaging network connectivity predicts substance abuse treatment completion. Biol. Psychiatry.: Cogn. Neurosci. Neuroimaging *3* (2), 141–149. https://doi.org/10.1016/j.bpsc.2017.07.003.

Sun, H., Jiang, R., Qi, S., Narr, K.L., Wade, B.S., Upston, J., Espinoza, R., Jones, T., Calhoun, V.D., Abbott, C.C., Sui, J., 2020. Preliminary prediction of individual response to electroconvulsive therapy using whole-brain functional magnetic resonance imaging data. NeuroImage: Clin. 26 https://doi.org/10.1016/j.nicl.2019.102080. Article 102080.

Taylor, J.J., Kurt, H.G., Anand, A., 2021. Resting state functional connectivity biomarkers of treatment response in mood disorders: a review. Front. Psychiatry 12. https://doi.org/10.3389/fpsyt.2021.565136. Article 565136.

Terrin, N., Schmid, C.H., Lau, J., Olkin, I., 2003. Adjusting for publication bias in the presence of heterogeneity. Stat. Med. *22* (13), 2113–2126. https://doi.org/10.1002/sim.1461.

Thölke, P., Mantilla-Ramos, Y.-J., Abdelhedi, H., Maschke, C., Dehgan, A., Harel, Y., Kemtur, A., Mekki Berrada, L., Sahraoui, M., Young, T., Bellemare Pépin, A., El Khantour, C., Landry, M., Pascarella, A., Hadid, V., Combrisson, E., O'Byrne, J., Jerbi, K., 2023. Class imbalance should not throw you off balance: choosing the right classifiers and performance metrics for brain decoding with imbalanced data. NeuroImage 277. https://doi.org/10.1016/j.neuroimage.2023.120253. Article 120253.

Tian, S., Sun, Y., Shao, J., Zhang, S., Mo, Z., Liu, X., Wang, Q., Wang, L., Zhao, P., Chattun, M.R., Yao, Z., Si, T., Lu, Q., 2020. Predicting escitalopram monotherapy response in depression: the role of anterior cingulate cortex. Hum. Brain Mapp. 41 (5), 1249–1260. https://doi.org/10.1002/hbm.24872.

Varoquaux, G., 2018. Cross-validation failure: small sample sizes lead to large error bars. NeuroImage *180* (Part A)), 68–77. https://doi.org/10.1016/j.neuroimage.2017.06.061.

Viechtbauer, W., 2005. Bias and efficiency of meta-analytic variance estimators in the random-effects model. J. Educ. Behav. Stat. *30* (3), 261–293. https://doi.org/10.3102/10769986030003261.

Vieira, S., Liang, X., Guiomar, R., Mechelli, A., 2022. Can we predict who will benefit from cognitive-behavioural therapy? A systematic review and meta-analysis of machine learning studies. Clin. Psychol. Rev. 97 https://doi.org/10.1016/j.cpr.2022.102193. Article 102193.

van Waarde, J.A., Scholte, H.S., van Oudheusden, L.J.B., Verwey, B., Denys, D., van Wingen, G.A., 2015. A functional MRI marker may predict the outcome of electroconvulsive therapy in severe and treatment-resistant depression. Mol. Psychiatry *20* (5), 609–614. https://doi.org/10.1038/mp.2014.78.

Watts, D., Pulice, R.F., Reilly, J., Brunoni, A.R., Kapczinski, F., Passos, I.C., 2022. Predicting treatment response using EEG in major depressive disorder: a machine-learning meta-analysis. Transl. Psychiatry 12 (1). https://doi.org/10.1038/s41398-022-02064-z. Article 332.

Wergeland, G.J.H., Riise, E.N., Öst, L.-G., 2021. Cognitive behavior therapy for internalizing disorders in children and adolescents in routine clinical care: a systematic review and meta-analysis. Clin. Psychol. Rev. *83*, 101918 https://doi.org/10.1016/j.cpr.2020.101918.

Whitfield-Gabrieli, S., Ghosh, S.S., Nieto-Castanon, A., Saygin, Z., Doehrmann, O., Chai, X.J., Reynolds, G.O., Hofmann, S.G., Pollack, M.H., Gabrieli, J.D.E., 2016. Brain connectomics predict response to treatment in social anxiety disorder. Mol. Psychiatry 21 (5), 680–685. https://doi.org/10.1038/mp.2015.109.

Williams, L.M., 2017. Defining biotypes for depression and anxiety based on large-scale circuit dysfunction: a theoretical review of the evidence and future directions for clinical translation. Depress Anxiety 34 (1), 9–24. https://doi.org/10.1002/da.22556.

Wolff, R.F., Moons, K.G.M., Riley, R.D., Whiting, P.F., Westwood, M., Collins, G.S., Reitsma, J.B., Kleijnen, J., Mallett, S., 2019. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. Ann. Intern. Med. *170* (1), 51–58. https://doi.org/10.7326/M18-1376.

Wu, H., Liu, R., Zhou, J., Feng, L., Wang, Y., Chen, X., Zhang, Z., Cui, J., Zhou, Y., Wang, G., 2022. Prediction of remission among patients with a major depressive disorder based on the resting-state functional connectivity of emotion regulation networks. Transl. Psychiatry 12. https://doi.org/10.1038/s41398-022-02152-0. Article 391.

Wu, Y., Mo, J., Sui, L., Zhang, J., Hu, W., Zhang, C., Wang, Y., Liu, C., Zhao, B., Wang, X., Zhang, K., Xie, X., 2021. Deep brain stimulation in treatment-resistant depression: a systematic review and meta-analysis on efficacy and safety. Front. Neurosci. 15, Article 655412. doi:10.3389/fnins.2021.655412 .

Xia, M., Liu, J., Mechelli, A., Sun, X., Ma, Q., Wang, X., Wei, D., Chen, Y., Liu, B., Huang, C.-C., Zheng, Y., Wu, Y., Chen, T., Cheng, Y., Xu, X., Gong, Q., Si, T., Qiu, S., Lin, C.-P., He, Y., 2022. Connectome gradient dysfunction in major depression and its association with gene expression profiles and treatment outcomes. Mol. Psychiatry 27 (3), 1384–1393. https://doi.org/10.1038/s41380-022-01519-5.

Yarkoni, T., Westfall, J., 2017. Choosing prediction over explanation in psychology: lessons from machine learning. Perspect. Psychol. Sci.: A J. Assoc. Psychol. Sci. *12* (6), 1100–1122. https://doi.org/10.1177/1745691617693393.

Zeng, L.-L., Shen, H., Liu, L., Wang, L., Li, B., Fang, P., Zhou, Z., Li, Y., Hu, D., 2012. Identifying major depression using whole-brain functional connectivity: A multivariate pattern analysis. Brain 135 (5), 1498–1507. https://doi.org/10.1093/brain/aws059.

Zhutovsky, P., Thomas, R.M., Olff, M., van Rooij, S.J.H., Kennis, M., van Wingen, G.A., Geuze, E., 2019. Individual prediction of psychotherapy outcome in posttraumatic stress disorder using neuroimaging data. Transl. Psychiatry 9. https://doi.org/10.1038/s41398-019-0663-7. Article 326.

Zhutovsky, P., Zantvoord, J.B., Ensink, J.B.M., Op den Kelder, R., Lindauer, R.J.L., van Wingen, G.A., 2021. Individual prediction of trauma-focused psychotherapy response in youth with posttraumatic stress disorder using resting-state functional connectivity. NeuroImage: Clin. 32 https://doi.org/10.1016/j.nicl.2021.102898. Article 102898.